A framework for identifying genotypic information from clinical records: exploiting integrated ontology structures to transfer annotations between ICD codes and Gene Ontologies

Seyedsasan Hashemikhabir, Ran Xia, Yang Xiang, Sarath Chandra Janga

Abstract— Although some methods are proposed for automatic ontology generation, none of them address the issue of integrating large-scale heterogeneous biomedical ontologies. We propose a novel approach for integrating various types of ontologies efficiently and apply it to integrate International Classification of Diseases, Ninth Revision, Clinical Modification (ICD9CM) and Gene Ontologies (GO). This approach is one of the early attempts to quantify the associations among clinical terms (e.g. ICD9 codes) based on their corresponding genomic relationships. We reconstructed a merged tree for a partial set of GO and ICD9 codes and measured the performance of this tree in terms of associations' relevance by comparing them with two well-known disease-gene datasets (i.e. MalaCards and Disease Ontology). Furthermore, we compared the genomic-based ICD9 associations to temporal relationships between them from electronic health records. Our analysis shows promising associations supported by both comparisons suggesting a high reliability. We also manually analyzed several significant associations and found promising support from literature.

Index Terms— knowledge integration, clinical records, Gene Ontology

1 Introduction

AUTOMATIC ontology generation and integration are desirable in many applications and have been studied in the past decade. Although available methods for automatic ontology generation produce ontologies from a single type of data, such as gene networks [1], textual data [2], dictionary [3] and schemata [4], [5], they do not contribute to the integration of different types of ontologies which will bring innovative results on annotation/knowledge reuse and association studies. A number of studies have been focused on ontology integration [6], [7] and their medical domain applications[8]. The ontology integration methods are generally classified into three categories: 1) Manual set up of the integration rules [9]; 2) Heuristic approaches for automatic ontology integration[10], [11], [12], [13]; 3) Machine learning methods to assist the automatic ontology integrations [14]. These methods have a few major weaknesses including: 1) lack of efficient or systematic approaches to identify similarity between heterogeneous ontology concepts; 2) generally heuristic with no theoretical results to show the proposed integration approach is globally optimal; 3) they are not developed for integrating big ontologies or a large number of ontologies.

Our recent work on the UMLS (Unified Medical Language System) indexing [15] and UMLS mapping [16] allowed us to make necessary preparations for the heterogeneous ontology integration. The UMLS is a network of biomedical terms whose size and density exceeds the capacity of prevailing distance indexing methods for general graphs[17], [18]. However, by observing the scale-free structure of the UMLS network, we are able to index it by iteratively removing high degree vertices from the UMLS network and assign their labels to vertices within their k neighborhoods. We name such an approach k-Decentralized Labeling Scheme (kDLS) [15]. We demonstrated that we are able to use kDLS to efficiently evaluate the closeness between any two UMLS terms via the discovered paths between them. Later, we developed onGrid [15] to advance the kDLS method, by considering concept semantic types in the concept closeness measurement. The advantages of onGrid over kDLS were demonstrated in [15]. In order to use kDLS or onGrid to study the relationships between biomedical concepts, we also need an efficient tool to map biomedical concepts to UMLS terms. However, we found that available methods, MetaMap [19], cTAKES [20], and the UMLS Metathesaurus Browser, either fail to consider a biomedical concept as a whole, or fail to tolerate character variations in words. Thus, we developed a layered dynamic programming approach (LDPMap) [16] that is able to efficiently map a biomedical concept to a UMLS term by measuring the similarity between biomedical terms at both the

- *Seyedsasan Hashemikhabir and Sarath Chandra Janga are with Department of Biohealth Informatics, School of Informatics and Computing, Indiana University Purdue University, 719 Indiana Ave Ste 319, Walker Plaza Building, Indianapolis, Indiana 46202. E-mail: hashemis@iu.edu, scjanga@iupui.edu.*
- *Ran Xia, Yang Xiang are with Department of Biomedical Informatics, Ohio State University, 3190 Graves Hall, 333 West 10th Avenue, Columbus, Ohio 43210. E-mail: Ran.Xia@osumc.edu, yxiang@bmi.osu.edu.*

word and concept levels. As a result, given any two biomedical ontologies, we are able to use LDPMap (for non-UMLS ontologies) and onGrid[15] to generate a relationship matrix which provides 1) a systematic measurement to quantify the goodness for the ontology integration; 2) theoretical results to show that the proposed integration approach is globally optimal or close to optimal; 3) capability for integrating large volume of ontologies. We have previously demonstrated that these newly developed algorithms are efficient and provide optimal or near optimal solutions for integrating large ontology datasets [25]. Each entry in the relationship matrix is a value of closeness measurement between the corresponding two concepts.

Briefly, as described earlier [25], the basic ontology integration problem in our work can be formulated as follows. Given ontology tree structures T_A and T_B and a closeness based relationship matrix M_{AB}, integrated ontology tree structure should meet the following two basic criteria.

1) For any two vertices x and y in T_A (or T_B), the lowest common ancestor LCA T_A (x,y) (or LCA T_B (x,y)) is contained by LCA T_{AB} (x,y).
2) It holds that argmax T_{AB} $f(T_{AB})=\sum_{X\in V(T_{AB})} M_{AB}(X)$. Here $M_{AB}(X)$ is the entry value in the closeness matrix for the corresponding two vertices (one from T_A and the other from T_B) contained in the node X. $M_{AB}(X)=0$ if X, a node of T_{AB}, contains only one vertex from T_A or T_B.

We name $f(T_{AB})$ the cohesion function of the integrated ontology T_{AB} and its value is the overall cohesion score of integrating T_A and T_B into T_{AB}. Correspondingly, we define function $g(T_A, T_B)=max_{T_{AB}}(\sum_{v\in V(T_{AB})} M_{AB}(v))$ as the maximum cohesion function for integrating the ontologies T_A and T_B and its value is the maximum overall cohesion score (or, simply, maximum cohesion score). Consequently, we will assume that for any two given ontologies, their relationship/similarity matrix discussed in section 2.1 is available for calculating the cohesion score.

Our approach is innovative in three fundamental aspects: 1) We designed onGrid[15] and LDPMap[16] to efficiently generate similarity matrix between two sets of heterogeneous biomedical ontologies; 2) Given a similarity matrix between two sets of ontologies, we are able to identify an optimal solutionfor integrating the two ontologies; 3) Our methods can integrate big ontologies and can be extended to integrate a large number of ontologies. However, it is important to note that this work is a proof of concept for the original approach developed earlier and only a small fraction of GO dataset is integrated at this stage. In addition, our results demonstrate that it is possible to use the merged biomedical ontologies and derived annotations generated by our methods for biomedical knowledge discovery.

Each ontology has a hierarchical structure which is often recorded in paths. In the UMLS, the file "MRHIER.RRF" records each path from an ontology term to its root. Hence, we can build an ontology tree (sometimes it is a forest) upon these paths. Such a tree maintains the ancestry-descendent relationships among its concepts. In other words, a concept A is an ancestor of another concept B in a path, if and only if A is the ancestor of B in the ontology tree. When merging two ontologies, the basic constraint is that ancestor-descendent relationships are preserved. That is, for any two concepts A and B in an ontology tree, if A is the ancestor of B, then this is also true in the merged ontology tree. This is reasonable otherwise the results are illogical. Under such a basic constraint, our goal is to generate a merged ontology tree for ICD9 codes and Gene Ontology (GO) dataset that maximizes the closeness scores.

One of the fundamental problems in biomedical sciences is to link the phenotype with the genotype of an organism. However, most of the existing clinical systems only provide phenotypic information i.e, disease or clinical manifestation of a patient's ill health with very limited genotypic cause being provided. One approach to address this fundamental gap in our knowledge and a step closer towards personalized medicine is to map clinical manifestations on to the underlying tissue-specific genotypes. In an effort to address this gap, we have attempted to use the proposed ontology integration framework using UMLS as the backbone, to transfer GO annotations for the human genome to ICD9 annotations, thereby facilitating the mapping of disease codes and clinical manifestations to the underlying genes, pathways and processes. We anticipate that with improvements in annotations for the human genome at the molecular and cellular level, it should be possible to accurately map the gene pool contributing to a clinical manifestation in an automated fashion by deploying frameworks such as those proposed in this study.

2 Materials and Methods

2.1 Systematic method for heterogeneous ontology integration and evaluation

Ontologies are important knowledge entities existing in many areas of biomedical research. The UMLS contains many popular ontologies such as Gene Ontology (GO), International Classification of Diseases, Ninth Revision, Clinical Modification (ICD9CM), and Medical Dictionary for Regulatory Activities Terminology (MedDRA). These ontologies have been widely used in many biomedical applications. For example, GO has been used in designing TOPPGENE [21], a popular gene enrichment validation method, and ICD9CM has been used in many electronic health record (EHR) systems. However, despite the rich information provided by existing ontologies, they are only focused on the narrow knowledge domain of their own. The GO is annotated with gene information while the ICD9CM is annotated with disease and clinical information. There is no information sharing between these ontologies.

A significant amount of work in biomedical research is to identify the associations between different entities, such as identifying disease genes [22], prioritizing disease genes [23], and mapping phenotypes to genotypes [24]. These works will not benefit from a single ontology. If we can reuse (or map) the annotation of one ontology to another, it will significantly contribute to the study of various associations. In addition, not all ontologies are as well-studied as the most popular ontologies like GO. The ability to use the

annotation of one ontology to another will significantly facilitate our understanding on less studied ontologies. However, ontologies are often very large. GO, for example, contains over 300k terms. Manually integrating two ontologies is often too costly to be feasible, not to mention in real applications we may need to integrate many pairs of ontologies.

Available methods on ontology integration have several major weaknesses for effectively integrating biomedical ontologies in a large scale, as discussed earlier. To address this challenging problem, we designed a systematic approach on efficiently integrating large scale ontologies[25] . Our approach is built upon our past work on indexing the UMLS by decentralization labeling schema [15] and mapping concepts to the UMLS terms[16]. Using LDPMap and onGrid[15], we are able to map biomedical concepts to the UMLS CUIs (i.e, Concept Unique Identifiers), and generate relationship/similarity matrix between two sets of biomedical terms based on the discovered UMLS paths (transitive relationships) between them. The discovered UMLS paths between the concepts are measured as closeness values in this network, in order to construct a relationship/similarity matrix between the corresponding concepts. The workflow for generating similarity matrix is illustrated in the left part of Figure 1. If a biomedical ontology is chosen from the UMLS, then the LDPMap process can be skipped because each ontology term already has a CUI. With the generated similarity matrix, we are able to build an integrated ontology with a score of merging the two ontologies, as illustrated in the right panel of Figure 1. To get a maximum merging score, we can simply choose the maximum values in the similarity matrix and build the merged ontology correspondingly. However, by doing so we may get a merged ontology which violates the basic logic in the original ontology, such as the reverse of ancestor-descendent relationship. Therefore, our method balances the effort between maximizing the merging score and maintaining the basic hierarchical logic in the original ontologies. We realized this by a bottom-up approach which iteratively calculates the merge score for terms in the two ontologies in the reverse topological order. For example, when we calculate the merge score for term1a and term2a, the merge scores must have already been calculated between all terms with topological orders greater than term 1a in ontology 1 (i.e., term1b, term1c, term1d, term1e, term1f), and all terms with topological orders greater than term 1b in ontology 2 (i.e., term2b, term2c, term2d, term2e, term2f). As illustrated in Figure 1, before an entry in the merge score matrix is calculated, all entries up and left to this entry should have been calculated. However, the order of calculation is not unique. It is valid as long as it follows the reverse topological order. In the following text, we elaborate the major steps for integrating two given ontologies assuming their similarity matrix is available for identifying the ontology integration.

Step 1: Sort terms in ontology 1 and terms in ontology 2 in the reverse topological order;
Step 2: Initialize an empty merge score matrix whose row and column correspond to the sorted terms in ontology 1 and ontology 2;
Step 3: Pick up a pair of terms (term1x, term2y) in the sorted order;
Step 4: Calculate the maximal matching score MM between the children of term1x and the children of term2y;
Step 5: Let score(term1x, term2y)=max(MM+sim(term1x, term2y), maxi(score(i, term2y)), maxj(score(term1x, j)); where i and j are the indices of the terms upper or left of term1x and term2y respectively in the matrix.
Step 6: Allow term1x and term2y merge if MM+sim(term1x, term2y) is greater than maxi(score(i, term2y)) and maxj(score(term1x, j);
Step 7: Repeat from Step 3 if (term1x, term2y) is not the last pair in the sorted order;
Step 8: Output the integrated ontology.

Step 4 is a main process in the ontology integration. It determines the merging of children of term1x and term 2y for a maximal merge score MM. This can be achieved by iteratively selecting two child terms with maximal merge score, a process of matching child terms. If MM+sim(term1x, term2y) is greater than any scores in the score matrix upper or left to (term1x, term2y), then it suggests that merging term1x and term2y is the best choice in integrating the subtree rooted at term1x and the subtree rooted at term2y. Otherwise, it implies either merging term1x with a descendant term of term2y, or merging term1y with a descendant term of term1x, is the best choice in integrating the subtree rooted at term1x and the subtree rooted at term2y. Since we iteratively get the best choice for integrating each pair of subtrees, we eventually get the best choice for integrating ontology 1 and ontology 2. We found such a bottom-up approach works very effectively in integrating biomedical ontologies. However, the method is also limited by the memory size as we need to build matrices with sizes proportional to the product of the two ontology sizes.

In this study, we employed the above ontology integration framework to merge two different hierarchical annotation systems namely International Classification of Diseases, Ninth Revision (ICD9) codes and Gene Ontology (GO). ICD9CM (International Classification of Diseases, 9th edition, Clinical Modification) is a set of codes used by physicians, hospitals, and allied health workers to indicate diagnosis for all patient encounters.

In particular, we have implemented these basic algorithms in C++ and applied it on GO and ICD9CM, two popular ontologies in the UMLS. Our current implementationis limited by the memory size and is currently not able to handle extremely large ontologies. For instance, to complete the integration of the full GO and ICD9CM ontologies, it will take months to compute on a large memory machine. We plan to speed up the process in the future by developing distributed algorithms that work in a cloud computing environment. In this study, we used a subset of GO (a subtree from its root to vertices at depth 7) which includes 39,800 GO terms (out of a total of 333,507 terms) and the full set of ICD9CM which includes 22,401 terms.

2.2 Evaluation of the integration of ICD9 and GO to study the efficiency and effectiveness of the merged ontology and derived annotations

While it is not possible to obtain absolute gold standards (especially for disease annotations) to evaluate the efficiency and effectiveness of the merge process, due to various reasons including the semi-automatic curation efforts commonly employed, disease specific biases inherent in the databases, and lack of negative annotations for a disease etc., employing multiple datasets for evaluation of the annotations can be considered as a robust means to verify the quality and effectiveness of the merging process. Under this notion, we constructed the ICD9 to gene associations in the human genome by using the ICD9->GO relations derived, by employing our merging algorithm as a means to identify the genes contributing to the disease/phenotype with a given ICD9 code. Such predicted ICD9 to gene annotations come from transferring the gene annotations to ICD9 codes via GO IDs based on the integrated ontology on the UMLS. This data was used for evaluation of the merging process by studying enrichments comparing with the disease to gene annotations currently available from Disease Ontology (DO)[28] and MalaCards[29] databases, as they form two domain specific high-throughput datasets currently available for studying disease to gene annotations. To perform enrichments, we computed the overlap between the annotated ICD9 to gene associations from these resources and those discovered by our approach to calculate the hyper geometric probability of the overlap. P-values and corresponding significance values computed from them are used as statistical measures of our ability to reproduce the known associations. In particular, we performed the enrichments between our mapping for each ICD9 to gene annotations and these database annotations for the corresponding diseases to genes by calculating the hypergeometric probability of overlap. Similar statistical approach was adopted for computing enrichments when comparing predicted ICD9-ICD9 associations with the temporal ICD9-ICD9 associations data used in Section 3.3. Since these resources do not have true negative information, performing enrichments is a better approach as opposed to Receiver Operating Characteristics (ROC) analysis.

3 Results

3.1 Benchmarking the annotations resulting from integrated ontology with external databases

Figure 2 shows a sample of the integrated ontology with the relationships among ICD9 and GO annotations presented as ancestry and sibling associations along with all the shortest paths showing other UMLS ids contributing to the merge in this sub-tree. The generated network is easily interpretable and it can uncover the knowledge associated between ICD9 codes and their corresponding GO annotation terms. The network is modeled as directed network where a flow of information is originating from an ICD9 term and ends in a GO term. For example:

C0023448:Lymphoid leukemia → C1512385:Hematopoietic and Lymphoid Cell → C0007634:Cell → C1156236:Inhibition of Cell Proliferation → C1655747:negative regulation of cellular process

One potential interpretation of this shortest path between C0023448 (ICD9 code) and C1655747 (GO term) can demonstrate that "Lymphoid leukemia is originated from Hematopoietic Lymphoid cell and is associated with cell proliferation which is a negative regulation process." In another example:

C0153436:Malignant neoplasm of sigmoid colon → C0007102:Malignant tumor of colon → C0385939:SMAD4 protein, human → C0007586:Cell Cycle → C1659594:positive regulation of cell cycle

Similarly, the extracted path between C0153436 (ICD9 code) and C1659594 (GO term) which was present in the integrated ontology implies that "Malignant neoplasm of sigmoid colon is a tumor associated with malfunction of SMAD4 protein which consequently results in the positive regulation of cell cycle."

Previous examples demonstrated potential novel associations that can be extracted using this framework. Although verifying the associations requires further analysis, we verified the sample associations with literature and found experimental support. For example, Smurf2 [26] gene is known to be associated with tumor repressing mechanism and it is reported to influence Hodgkin's disease and was found to be annotated with various components of global cellular regulation in GO (e.g. death, differentiation, aging, adhesion). In another example, ST13 [27] is shown to be inhibiting colorectal cancer in cell lines that are directly associated with Malignant neoplasm of colon and Leukemia diseases. It is also reported that ST13[27] is associated with multiple global cellular regulatory pathways such as growth, morphogenesis and proliferation and our integrated ICD9 and GO annotation tree supports the mapping between these neoplasmic states and the cell regulation processes.

To further evaluate the quality of our disease to gene annotations, we integrated the GO annotations for humans (obtained from www.geneontology.org) with ICD9 codes, to construct a confident set of ICD9 to gene annotations for 220 ICD9 codes with varying number of genes associated. We then obtained disease to gene associations for the same set of ICD9 codes from the MalaCards[29] and Disease Ontology[28] datasets which are semi-automatically curated resources with disease-genotype information. MalaCards integrates various disease annotations (e.g. associated genes, therapeutics, anatomical context, disease classifications etc.) for nearly 10000 reported diseases in the literature. It contains more than 8100 diseases that are associated with 38000 non-unique ICD9 codes. Similarly, Disease Ontology (DO) datasets integrate several ontologies such as Medical Subject Headings (MeSH), ICD, Systematic Nomenclature of Medicine (SNOMED) and Online Mendelian Inheritance in Man (OMIM). It covers the annotations for 8700 diseases associated with 8300 ICD9 codes. The num-

ber of associated ICD9 and Genes are much lower in comparison to the MalaCards dataset. In order to increase the ICD9 association coverage in DO, we added the missing ICD9 codes from MalaCards to DO for the disease with same names. The improved DO dataset contained 27,500 non-unique ICD9 codes. We hypothesized that the confident ICD9 codes and their associated genes resulting from our approach should have a significant overlap with their corresponding diseases in both MalaCards and DO datasets. We removed the diseases from both MalaCards and DO that are associated with less than five genes. The remaining sets comprised of 2990 and 905 diseases for MalaCards and DO respectively. We then compared the extent of overlap in the gene sets between every possible ICD9 code and its corresponding disease to gene annotations in MalaCards and DO respectively. Since the merge between ICD9 and GO is not a complete merge of the entire ontologies, we believe studying the enrichment of disease-gene annotations in a disease centric manner using the known annotation resources like MalaCards and DO is a suitable approach to test the quality of our integration. While we don't intend to suggest that this is the only way to test the quality of our ICD9 to gene annotations, to our knowledge this is an ideal validation given the limited manually curated data in this domain. Out of a set of 112 diseases (50% of the total initial set of 220) which had at least 10 genes annotated in our final dataset, we found 27 and 17 diseases (Tables 1 and 2) to be significantly enriched (Hypergeometric) based on the associated genes documented in the MalaCards and Disease Ontology resources respectively. Our results suggest that the proposed approach is able to associate the correct GO to its corresponding ICD term as is evident from the high extent of overlap of gene sets based on the available annotations for these ICD9 codes in MalaCards and Disease Ontology resources. While our approach is not able to recover all the annotated genes based on our integration, it is possible that the partial integration of the ontologies is contributing to the poor coverage. Although MalaCards takes advantage of richer and more accurate annotation in comparison to Disease Ontology, we found 10 ICD9 codes (diseases) to be significant in our comparison with both the datasets ($p< 0.01$, Hypergeometric test) (Figure 3). These observations show that the genes under a given ICD9 code using our approach are usually enriched for the same disease using different disease annotation resources currently available. For instance, we found ICD9 codes associated with leukemia, hepatitis, asthma, colitis and colon cancer to be abundant in the diseases which exhibited significant overlap between our predictions and current annotations. It is possible that our observed prevalence of some of these diseases might reflect a bias in current annotations for certain diseases or due to the incomplete merge of ICD9-GO accomplished in this study and not necessarily due to the better performance of the integrated ontology on certain diseases.

Although our results show that for about 25% of the ICD9 codes with considerable number of genes annotated in our merged tree, our performance compares well to MalaCards and DO datasets, we also found that a number of ICD9 codes didn't show a significant overlap of genes. While various possibilities exist including 1) the fact that our merge is not complete (only a fraction of GO tree was merged with ICD9), 2) certain diseases might be more commonly studied and hence are likely to have higher number of annotations which may not all be true associations in existing databases and 3) finally the number of genes annotated for a given ICD9 code can also play a role given the incomplete disease annotations present in the databases. To evaluate the third scenario, we compared the number of genes annotated for highly significant ($p< 0.01$, Hypergeometric test) and non-significant (same number of diseases with high p-values) ICD9 codes in this comparison with our disease datasets (Figure 4). It is evident in all the three resources that higher the significance for overlap the higher was the number of genes, suggesting that lower overlap can be mainly due to incompleteness or fewer annotations in either of the two data resources being compared. These observations encourage the need to improve the coverage of the number of annotations by possibly expanding the GO levels to identify more disease to gene associations in our mapping framework.

3.2 Uncovering associations between diseases using the gene annotations to ICD9 codes

In order to further evaluate the quality of our associations, we compared the extent of overlap in genes between every pair of ICD9 codes in our mapped dataset, for a set of diseases which also exhibited highly significant enrichment for corresponding disease annotations in MalaCards and DO respectively. This resulted in association networks between ICD9 codes (Figure 5). We identified 34 and 19 significant associations (Hypergeometric) among the ICD9 codes for MalaCards (Figure 5A) and DO (Figure 5B) overlapping disease sets respectively. Although the generated networks are from independent annotation datasets, interestingly, there is a notable overlap between the associations of the two networks (i.e 9 associations) which indicates that our ontology integration method is one of the promising efforts in merging the annotations from various ontologies. Moreover, the associations also suggest the potential genomic linkages uncovered from our integration between diseases, like hodgkin's sarcoma and malignant neoplasm of colon[30] , malignant neoplasm of larynx and hemangioma of unspecified site[31], [32] as well as liver necrosis and obstruction of bile duct [33], [34], [35]. These results suggest that our integrated ontology framework can be used to not only transfer annotations from one level to the other but can also facilitate comparison and understanding of relationships between the annotated groups at higher levels. Thus, we believe the extension of our proposed integration framework would enable the discovery of novel disease-disease associations based on the sharing of genotypic information.

3.3 Comparison of genomic based ICD9 associations with temporal relationships based on electronic medical records

We showed that our results have a high overlap with currently well-known annotations linking diseases with

genes, however ICD9 codes are generally assigned based on clinical indications of diseases and it is possible to argue that diseases which are functionally associated based on sharing of genes may not be of clinical relevance. Hence, we asked whether ICD9 codes which were found to be associated due to sharing of genes could also have clinical relevance. In particular, we explored the overlap of genomic-based associations between ICD9 codes with those identified based on their linkages calculated by temporal relationships in electronic medical records. A recent study [36] from University of Michigan calculated the temporal relationships among the ICD9 codes based on the electronic medical records (EMRs). They extracted the pairwise associations among ICD9 codes employing X2 test and temporal data from the time stamps of the ICD9 codes and reported a total of 400000 top associations among the occurring ICD9 codes in the EMRs. We used this data to question if these epidemiological relationships between ICD9 codes can be captured based on our genotypic evidence i.e, sharing of annotated genes between ICD9 codes from the integrated ontology - as an aternate means of studying the quality of our integration. Since we did our analysis with a sample set of 220 ICD9 codes, we extracted the set of all associations among the same set from the temporal relationship dataset and we were able to retrieve 62 temporal associations. We assumed that ICD9-ICD9 associations reported based on temporal relationships are indeed the truly reported ones. We examined the quality of our predictions by comparing the overlap with their predictions. We calculated significance of overlap of two sets by assuming given global number of possible ICD9-ICD9 associations, whether the observed overlap is by chance. We calculated the significance value using hypergeometric distribution for the overlap of 45 MalaCards and DO associations between our result and the assumed gold standard. We found 15 significantly shared associations between our data and the temporal associations reported between ICD9 codes in the U. Michigan dataset (P-value<1e-32, Hypergeometric) (Figure 6 and Table 3).

To further test whether the predicted ICD9 to gene associations can be used to identify higher order literature quality interactions between diseases, we surveyed literature in support of these interactions. We found a strong support in the literature for several overlapping associations. For instance, Advair is one of the primary medicines for treating the patients with Asthma (C0004096), however, it is reported that oral candidiasis (C0006840) is one of the side-effect of this treatment [37]. A study [38] reported a 45-years old woman with increasing retrosternal tenderness as well as esophagitis (C0014868) that had benign tumor classified as cavernous haemangioma (C0018916). Another study[39] constructed a murine experimental model of bile duct obstruction that facilitates the controlled observations of the acute and subacute phases of cholestasis suggesting a strong association between acute and subacute necrosis of liver (C0001308) and obstruction of bile duct (C0008370). Several diseases might share underlying genotypes however our knowledge about their relationships is far from limited and these results supported by literature evidence strengthen our approaches ability to uncover such diseae-disease associations through automated mining techniques. More generally, these studies support the relevance of the genomic-based ICD9 associations identified using our framework in a translational clinical context.

4 Conclusion and Discussion

We proposed a novel approach for merging the annotations from multiple ontologies and applied it for integrating ICD9 and GO ontologies as a proof of concept. We showed that the associations obtained from the integrated tree are significantly enriched for annotations in comparison with MalaCards and Disease Ontology datasets. Although our integrated ontologies revealed promising associations with the genes, number of the associated genes plays an important role in our results. Therefore, it might be interesting to see how the impact of adding various levels of GO annotations would contribute to our integration framework as increasing the levels might increase the coverage of the annotations. To address this, we plan to develop a distributed algorithm that works in a cloud computing environment, and divide the ontology integration into sub tasks each with less resource requirement. Moreover, the quality of genes associated with each ontology can be measured by use of biological networks such Protein-Protein Interaction (PPI) networks. This potential layer of verification might help to eliminate the genes that are associated due to wrong annotations. Improving the coverage of our integration framework to build a global ICD9 to GO mapping together with a deeper understanding of the contribution of merging at different levels of the GO tree can provide valuable insights into the factors contributing to an efficient merge process. Future work in this direction will also enable the improvements in automatic annotations of poorly studied ontologies as well as improve our understanding on how various ontologies are interrelated.

Since our proposed framework for ontology integration in this study, depends on the pairwise similarity between every pair of UMLS terms (after mapping both ontologies onto the UMLS IDs) by measuring the path lengths in the merged hierarchy maintained tree and then optimizing the distance matrix, this approach can also work on ontology structures with multiple hierarchies. However, if the interest is to merge each hierarchy (say GO process only) of a given ontology independently with target ontology as opposed to doing a combined merge - this is also possible. For instance, if one of the hierarchies is more completely annotated it might be a preferred option to choose to merge only a specific branch. Hence, given the generic nature of integration using distance matrices after merging two different ontologies over UMLS, this framework can be a promising approach for integrating a wide range of ontological structures at varying levels of resolution for knowledge discovery by annotation repurposing.

Acknowledgment

SCJ acknowledges support from the School of Informatics and Computing at Indiana University Purdue University

Indianapolis (IUPUI) in the form of start-up funds. The authors would also like to thank members of the Janga Lab for providing helpful feedback in the course of this study. We would like to thank Dr. David Hanauer for sharing their results on the temporal relationships among the ICD9 codes based on the electronic medical records.

REFERENCES

[1] J. Dutkowski, M. Kramer, M. A. Surma *et al.*, "A gene ontology inferred from molecular networks," *Nat Biotech,* vol. 31, no. 1, pp. 38-45, 01//print, 2013.

[2] R. Navigli, P. Velardi, and A. Gangemi, "Ontology Learning and Its Application to Automated Terminology Translation," *IEEE Intelligent Systems,* vol. 18, no. 1, pp. 22-31, 2003.

[3] J. Jannink, "Thesaurus Entry Extraction from an On-line Dictionary," 1999.

[4] C. Papatheodorou, A. Vassiliou, and B. Simon, "Discovery of Ontologies for Learning Resources using Word-based Clustering," in World Conference on Educational Multimedia, Hypermedia and Telecommunications 2002, Denver, Colorado, USA, 2002, pp. 1523-1528.

[5] D. L. Rubin, M. Hewett, D. E. Oliver *et al.*, "Automating data acquisition into ontologies from pharmacogenetics relational data sources using declarative object definitions and XML," *Pac Symp Biocomput*, pp. 88-99, 2002.

[6] N. F. Noy, "Semantic integration: a survey of ontology-based approaches," *SIGMOD Rec.,* vol. 33, no. 4, pp. 65-70, 2004.

[7] S. Abels, L. Haak, and A. Hahn, "Identification of common methods used for ontology integration tasks," in Proceedings of the first international workshop on Interoperability of heterogeneous information systems, Bremen, Germany, 2005, pp. 75-78.

[8] A. G. a. D. M. P. a. G. Steve, "Ontology Integration: Experiences with Medical Terminologies," pp. 163--178, 1998.

[9] H. S. Pinto, Jo\, \#227 *et al.*, "A methodology for ontology integration," in Proceedings of the 1st international conference on Knowledge capture, Victoria, British Columbia, Canada, 2001, pp. 131-138.

[10] J. Xie, F. Liu, and S.-U. Guan, "Tree-structure Based Ontology Integration," *Journal of Information Science,* vol. 37, no. 6, pp. 594-613, December 1, 2011, 2011.

[11] D. C. a. G. D. G. a. M. Lenzerini, "A Framework for Ontology Integration," *IOS Press*, pp. 303--316, 2001.

[12] O. Udrea, L. Getoor, Ren *et al.*, "Leveraging data and structure in ontology integration," in Proceedings of the 2007 ACM SIGMOD international conference on Management of data, Beijing, China, 2007, pp. 449-460.

[13] N. F. N. a. M. A. Musen, "SMART: Automated Support for Ontology Merging and Alignment," 1999.

[14] A. Doan, J. Madhavan, P. Domingos *et al.*, "Learning to map between ontologies on the semantic web," in Proceedings of the 11th international conference on World Wide Web, Honolulu, Hawaii, USA, 2002, pp. 662-673.

[15] A. Albin, X. Ji, T. B. Borlawsky *et al.*, "Enabling Online Studies of Conceptual Relationships Between Medical Terms: Developing an Efficient Web Platform," *JMIR Medical Informatics,* vol. 2, no. 2, pp. e23, 2014.

[16] A. L. K. Ren, A. Mukhopadhyay, R. Machiraju, K. Huang, Y. Xiang, "Effectively processing medical term queries on the UMLS Metathesaurus by Layered Dynamic Programming," *BMC Medical Genomics*, 2013.

[17] E. Cohen, E. Halperin, H. Kaplan *et al.*, "Reachability and distance queries via 2-hop labels," in Proceedings of the thirteenth annual ACM-SIAM symposium on Discrete algorithms, San Francisco, California, 2002, pp. 937-946.

[18] J. Cheng, and J. X. Yu, "On-line exact shortest distance query processing," in Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology, Saint Petersburg, Russia, 2009, pp. 481-492.

[19] A. R. Aronson, "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program," *Proc AMIA Symp*, pp. 17-21, 2001.

[20] G. K. Savova, J. J. Masanz, P. V. Ogren *et al.*, "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications," *J Am Med Inform Assoc,* vol. 17, no. 5, pp. 507-13, Sep-Oct, 2010.

[21] J. Chen, E. E. Bardes, B. J. Aronow *et al.*, "ToppGene Suite for gene list enrichment analysis and candidate gene prioritization," *Nucleic Acids Res,* vol. 37, no. Web Server issue, pp. W305-11, Jul, 2009.

[22] J. Zhang, K. Lu, Y. Xiang *et al.*, "Weighted frequent gene co-expression network mining to identify genes involved in genome stability," *PLoS Comput Biol,* vol. 8, no. 8, pp. e1002656, 2012.

[23] Y. Xiang, P. R. Payne, and K. Huang, "Transactional database transformation and its application in prioritizing human disease genes," *IEEE/ACM Trans Comput Biol Bioinform,* vol. 9, no. 1, pp. 294-304, Jan-Feb, 2012.

[24] L. A. Hindorff, P. Sethupathy, H. A. Junkins *et al.*, "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits," *Proc Natl Acad Sci U S A,* vol. 106, no. 23, pp. 9362-7, Jun 9, 2009.

[25] Y. Xiang, and S. C. Janga, "Building Integrated Ontological Knowledge Structures with Efficient Approximation Algorithms," *BioMed Research International*.

[26] M. Blank, Y. Tang, M. Yamashita *et al.*, "A tumor suppressor function of Smurf2 associated with controlling chromatin landscape and genome stability through RNF20," *Nat Med,* vol. 18, no. 2, pp. 227-34, Feb, 2012.

[27] R. Bai, Z. Shi, J. W. Zhang *et al.*, "ST13, a proliferation regulator, inhibits growth and migration of colorectal cancer cell lines," *J Zhejiang Univ Sci B,* vol. 13, no. 11, pp. 884-93, Nov, 2012.

[28] K. Peng, W. Xu, J. Zheng *et al.*, "The Disease and Gene Annotations (DGA): an annotation resource for human

pp. D553-60, Jan, 2013.

[29] N. Rappaport, N. Nativ, G. Stelzer *et al.*, "MalaCards: an integrated compendium for diseases and their annotation," *Database (Oxford),* vol. 2013, pp. bat018, 2013.

[30] M. R. Kashi, L. Belayev, and A. Parker, "Primary extranodal Hodgkin lymphoma of the colon masquerading as new diagnosis of Crohn's disease," *Clin Gastroenterol Hepatol,* vol. 8, no. 10, pp. A20, Oct, 2010.

[31] J. Wu, M. Qian, and J. Zhou, "[Clinical analysis of granulomatous capillary hemangioma of the larynx]," *Lin Chung Er Bi Yan Hou Tou Jing Wai Ke Za Zhi,* vol. 26, no. 15, pp. 677-9, Aug, 2012.

[32] S. Akhtar, A. A. Shamim, S. Ghaffar *et al.*, "Adult laryngeal haemangioma; a rare entity," *J Pak Med Assoc,* vol. 62, no. 2, pp. 173-4, Feb, 2012.

[33] R. Heimann, "Factors producing liver cell necrosis in experimental obstruction of the common bile duct," *The Journal of Pathology and Bacteriology,* vol. 90, no. 2, pp. 479-485, 1965.

[34] M. Yang, A. Ramachandran, H. M. Yan *et al.*, "Osteopontin is an initial mediator of inflammation and liver injury during obstructive cholestasis after bile duct ligation in mice," *Toxicol Lett,* vol. 224, no. 2, pp. 186-95, Jan 13, 2014.

[35] B. L. Woolbright, D. J. Antoine, R. E. Jenkins *et al.*, "Plasma biomarkers of liver injury and inflammation demonstrate a lack of apoptosis during obstructive cholestasis in mice," *Toxicol Appl Pharmacol,* vol. 273, no. 3, pp. 524-31, Dec 15, 2013.

[36] D. A. Hanauer, and N. Ramakrishnan, "Modeling temporal relationships in large scale clinical associations," *J Am Med Inform Assoc,* vol. 20, no. 2, pp. 332-41, Mar-Apr, 2013.

[37] C. R. Pinto, N. R. Almeida, T. S. Marques *et al.*, "Local adverse effects associated with the use of inhaled corticosteroids in patients with moderate or severe asthma," *J Bras Pneumol,* vol. 39, no. 4, pp. 409-17, Jun-Aug, 2013.

[38] H. U. Jahn, B. Matthees, H. J. Wenisch *et al.*, "[Cavernous haemangioma in the Muscularis propria of the oesophagus and in the paraoesophageal tissue]," *Z Gastroenterol,* vol. 40, no. 6, pp. 413-8, Jun, 2002.

[39] I. B. Prado, M. H. dos Santos, F. P. Lopasso *et al.*, "Cholestasis in a murine experimental model: lesions include hepatocyte ischemic necrosis," *Rev Hosp Clin Fac Med Sao Paulo,* vol. 58, no. 1, pp. 27-32, Jan-Feb, 2003.

Seyedsasan Hashemikhabir. He completed his Bachelors degree in computer engineering from Urmia University, Iran. In 2012, he obtained his Masters degree in computer engineering from Middle East Technical University, Turkey. During his master studies, he studied the core bioinformatics courses and involved in many projects related to Systems biology with emphasis on biological networks. He is currently a research scholar in Janga Lab. His current research interests are Systems Biology with emphasis on gene function prediction and studying the properties of biological networks with the use of graph algorithms and data mining methods.

Ran Xia. He is a Master student in computer science and engineering from the Ohio State University, Columbus, U.S.

Yang Xiang. He received the PhD degree in computer science from Kent State University in 2009. In 2010, he received the computing innovation fellow (CIFellow) award from US National Science Foundation (NSF)/CRA/CCC. In 2012, He was appointed as a research assistant professor in the Department of Biomedical Informatics, The Ohio State University. He is a member of the IEEE.

Sarath Chandra Janga. He obtained his PhD from the MRC Laboratory of Molecular Biology & University of Cambridge in 2010. He is currently an Assistant Professor of Informatics in the Department of Biohealth Informatics at the School of Informatics and Computing, Indiana University Purdue University at Indianapolis and a faculty member of the Center for Computational Biology and Bioinformatics at the Indiana University School of Medicine. Sarath's research interests include understanding the design principles and constraints imposed on gene regulatory systems within the broader field of computational and systems biology his lab works on.

FIGURES

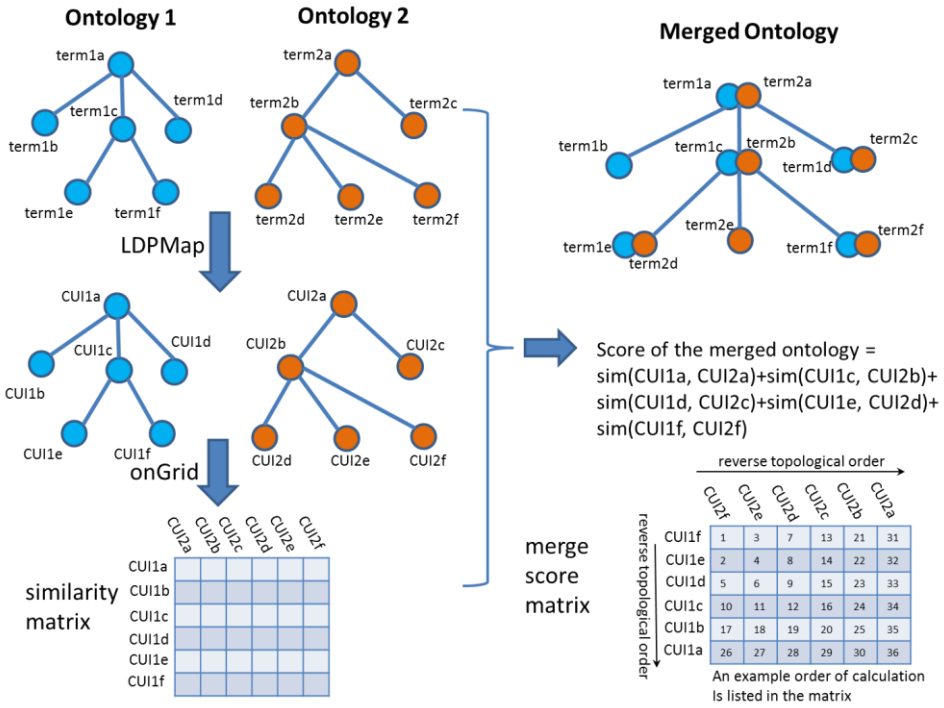

Fig. 1. Workflow showing the procedure for integrating two heterogeneous ontology trees. Panel to the left shows the mapping of a typical non-UMLS ontology onto UMLS CUIs using LDPmap (this step is not needed for a UMLS-based ontology) followed by using the kDLS approach for generating a similarity matrix between the two ontologies. Panel to the right shows the merged ontology which maintains the hierarchy of the original concepts as well as minimizes the score of the similarity matrix to generate a globally optimal merge.

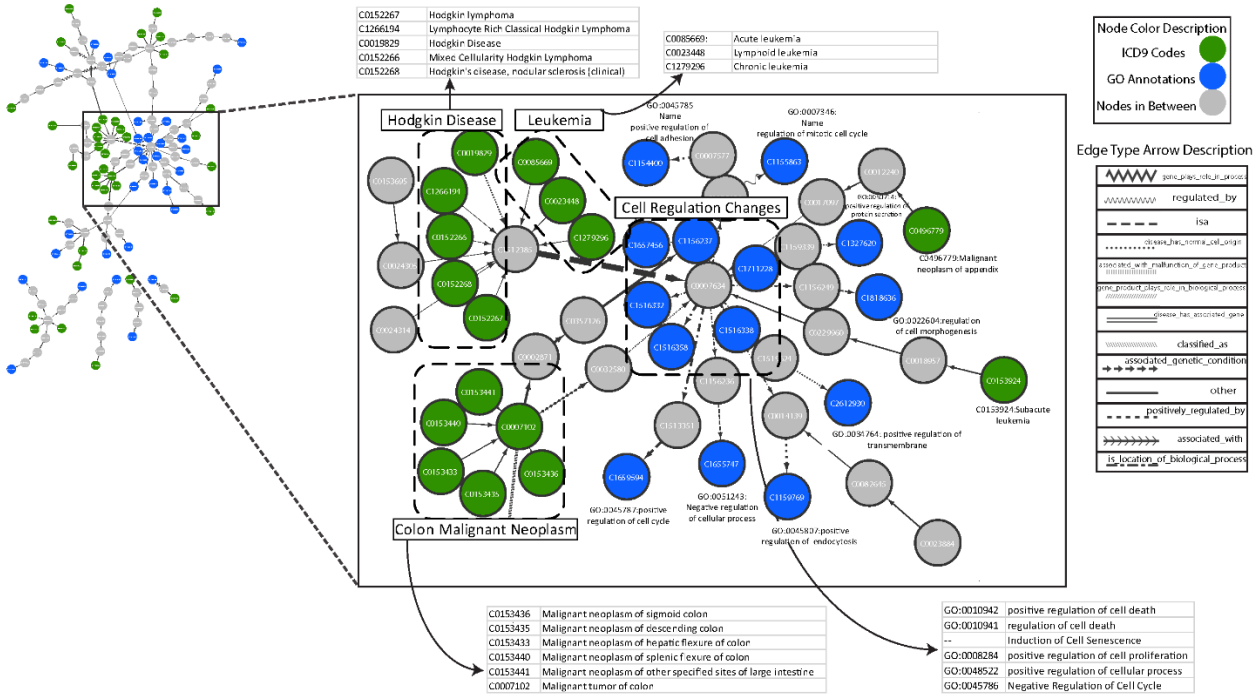

Fig. 2. Sub-tree from the merged ICD9 and GO tree showing the relationships and paths between various diseases (shown as ICD9 codes in green) and Gene Ontology terms (shown as GO terms in blue). This extracted tree highlights the various ICD9 annotations as CUIs, shows the intermediate connecting UMLS ids (CUIs) in grey as well as the edge types highlighted with a legend to the right. Other UMLS nodes (grey nodes) which facilitated this merge process in the form of shortest paths to connect these concepts are also shown.

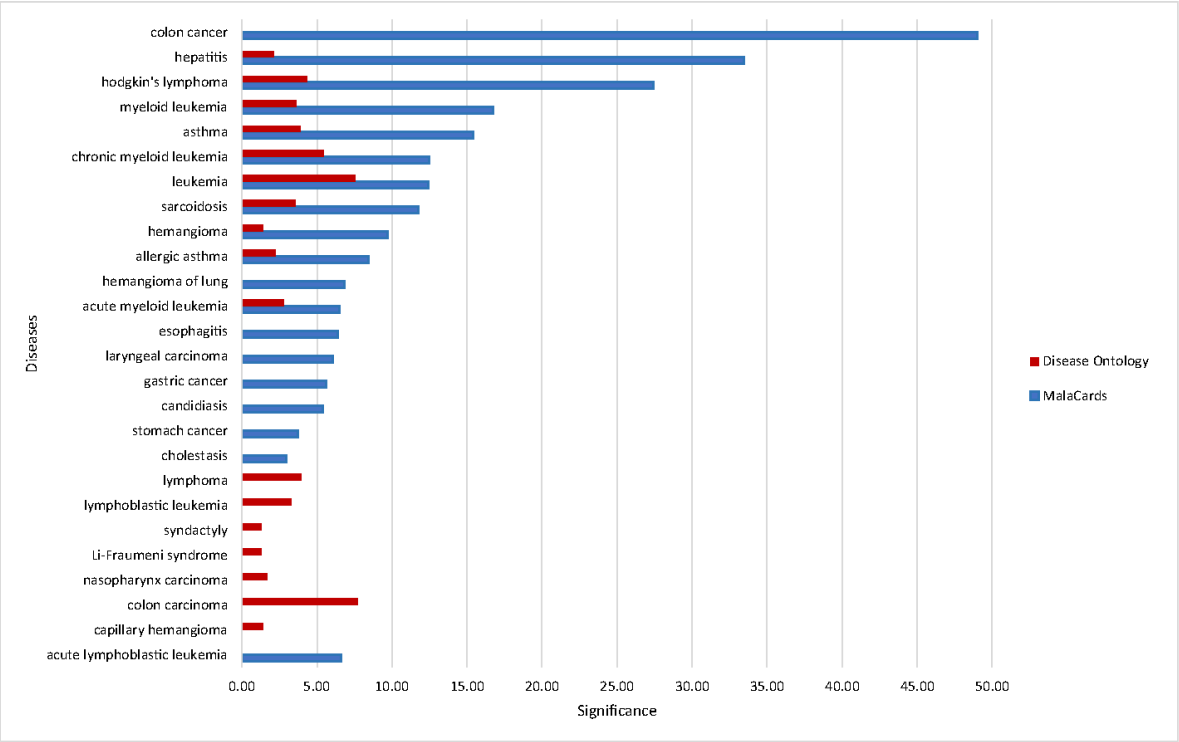

Fig. 3. Diseases with significant overlap in the number of genes between ICD9->GO mapping and MalaCards as well as DO annotations respectively. The figure shows the list of the diseases (based on ICD9 codes) that were found to be highly enriched. Each bar shows significance calculated as –log(p-value) of enrichment for a given ICD9/disease.

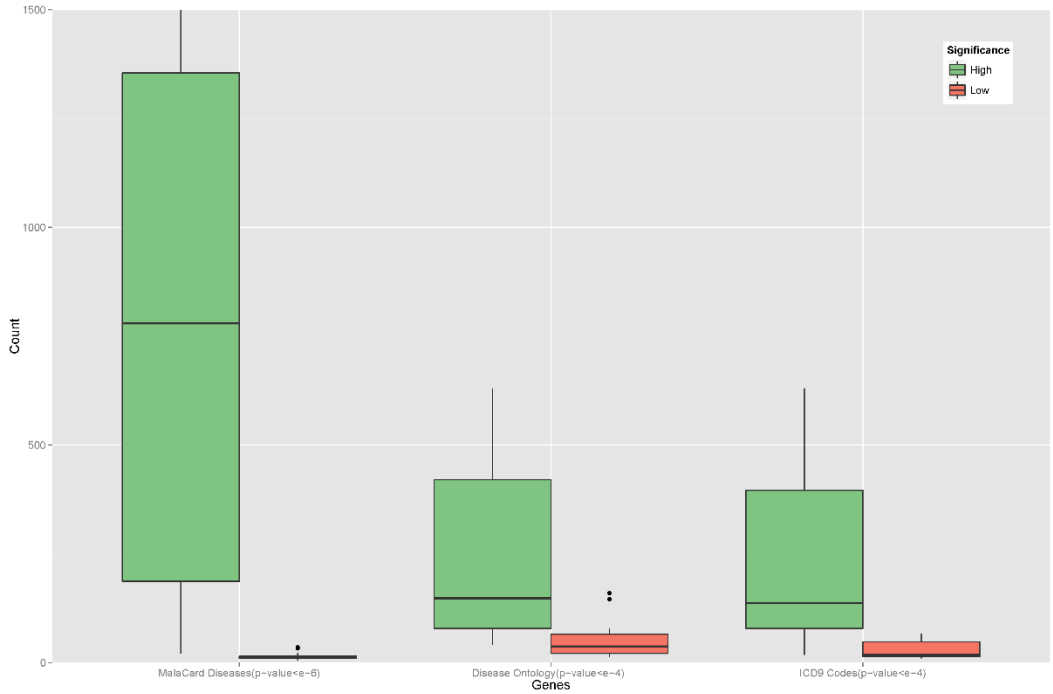

Fig. 4. In this analysis we hypothesized that the number of the genes associated with MalaCards, DO disease annotations and ICD9 codes are directly associated with significance of the observed results. In other words, we wanted to know the relation between the observed significance and the extent of annotation in the respective databases. To test this, we formed two groups, each containing the reported associations with highest and lowest significance and counted the number of genes reported for each pair of ICD9 code, MalaCards and DO diseases. We found that gene count was significantly higher in the high significance set than in the set with no significance. P-values based on Wilcoxon tests are shown under each comparison.

ICD Code	Description
C0007102	Malignant neoplasm of colon, unspecified
C0019829	Hodgkin's sarcoma
C0023418	Leukemia of unspecified cell type
C0004096	Asthma
C0023448	Lymphoid leukemia
C0018916	Hemangioma of unspecified site
C0006413	Burkitt's tumor or lymphoma
C0039075	Syndactyly
C0001308	Acute and subacute necrosis of liver
C0014868	Esophagitis, unspecified
C0007107	Malignant neoplasm of larynx, unspecified
C0024623	Malignant neoplasm of stomach, unspecified
C0006840	Candidiasis of unspecified site
C0008370	Obstruction of bile duct

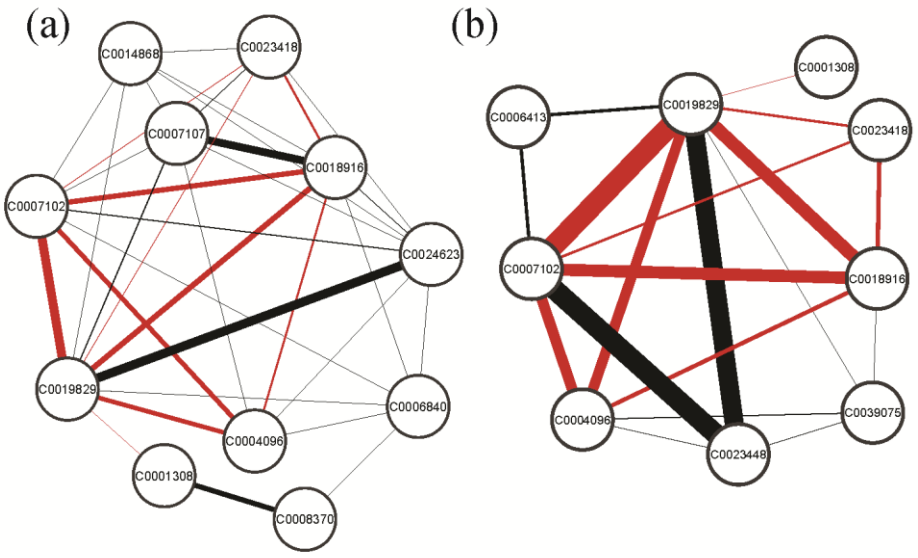

Fig. 5. Association networks for (a) MalaCards and (b) Disease Ontology showing the significance of association among the ICD9 Codes based on their gene pool overlap. In this analysis we calculated hypergeometric p-values for every pair of ICD9 Codes of interest using our ICD9 to gene mapping and the associations with p-value<0.05 are shown with significance values proportional to the thickness (measured as –log(p-value) of enrichment) of the edge.

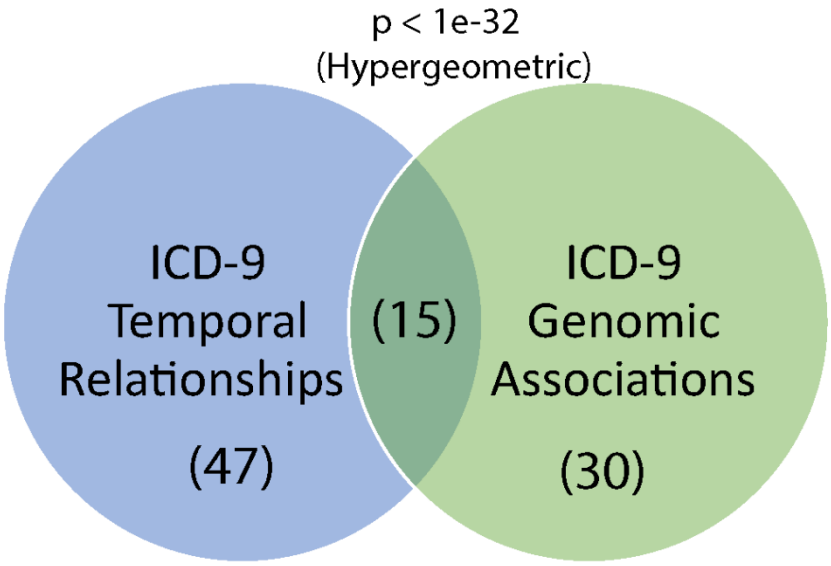

Fig. 6. Overlap of ICD9-ICD9 gene based associations with temporal relationships data obtained from Hanauer and Ramakrishnan (Hanauer and Ramakrishnan, 2013).

TABLES

Table 1. Highly significant pair of ICD9-MalaCard associations (p-value<0.01).

ICD9 Code (CUI)	ICD9 Description	Gene Count (ICD9)	MalaCard Disease Name	Gene Count(MalaCards)	Signifi-cance
C0007102	Malignant neoplasm of colon, unspecified	420	colon cancer	1491	9.48E-50
C0001308	Acute and subacute necrosis of liver	148	hepatitis	3143	3.42E-34
C0019829	Hodgkin's sarcoma	498	hodgkin's lymphoma	780	3.77E-28
C0023418	Leukemia of unspecified cell type	79	myeloid leukemia	1219	1.78E-17
C0004096	Asthma	630	asthma	901	3.94E-16
C0023418	Leukemia of unspecified cell type	79	chronic myeloid leukemia	427	3.27E-13
C0023418	Leukemia of unspecified cell type	79	leukemia	3411	3.75E-13
C0019829	Hodgkin's sarcoma	498	sarcoidosis	278	1.69E-12
C0018916	Hemangioma of unspecified site	372	hemangioma	163	1.98E-10
C0004096	Asthma	630	allergic asthma	138	3.43E-09
C0018916	Hemangioma of unspecified site	372	hemangioma of lung	21	1.48E-07
C0023418	Leukemia of unspecified cell type	79	acute lymphoblastic leukemia	783	2.31E-07
C0023418	Leukemia of unspecified cell type	79	acute myeloid leukemia	804	3.28E-07
C0014868	Esophagitis, unspecified	147	esophagitis	1503	4.00E-07
C0007107	Malignant neoplasm of larynx, unspecified	137	laryngeal carcinoma	177	9.32E-07
C0024623	Malignant neoplasm of stomach, unspecified	27	gastric cancer	2103	2.41E-06
C0006840	Candidiasis of unspecified site	18	candidiasis	125	4.36E-06
C0024623	Malignant neoplasm of stomach, unspecified	27	stomach cancer	197	1.91E-04
C0008370	Obstruction of bile duct	39	cholestasis	212	1.08E-03
C0023418	Leukemia of unspecified cell type	79	acute lymphocytic leukemia	112	1.43E-03
C0019360	Herpes zoster	34	herpes zoster	36	1.52E-03
C0023418	Leukemia of unspecified cell type	79	acute promyelocytic leukemia	231	3.12E-03
C0023418	Leukemia of unspecified cell type	79	li-fraumeni syndrome	20	3.44E-03
C0006413	Burkitt's tumor or lymphoma	108	burkitt's lymphoma	270	4.80E-03
C0018099	Gout, unspecified	58	gout	260	6.20E-03
C0009324	Ulcerative colitis, unspecified	142	ulcerative colitis	583	8.48E-03
C0023448	Lymphoid leukemia	41	lymphoblastic leukemia	872	8.63E-03

Table 2. Highly significant pair of ICD9-DO associations (p-value<0.05)

ICD9 Code (CUI)	ICD9 Description	Gene Count	DO Disease Name	Gene Count	Significance
C0023418	Leukemia of unspecified cell type	79	leukemia	982	2.84E-08
C0023418	Leukemia of unspecified cell type	79	chronic myeloid leukemia	172	3.64E-06
C0019829	Hodgkin's sarcoma	498	Hodgkin's lymphoma	66	4.73E-05
C0006413	Burkitt's tumor or lymphoma	108	lymphoma	561	1.11E-04
C0004096	Asthma	630	asthma	351	1.24E-04
C0023418	Leukemia of unspecified cell type	79	myeloid leukemia	367	2.48E-04
C0019829	Hodgkin's sarcoma	498	sarcoidosis	79	2.63E-04
C0023448	Lymphoid leukemia	41	lymphoblastic leukemia	206	5.41E-04
C0023418	Leukemia of unspecified cell type	79	acute myeloid leukemia	388	1.52E-03
C0004096	Asthma	630	allergic asthma	39	5.70E-03
C0001308	Acute and subacute necrosis of liver	148	hepatitis	423	7.14E-03
C0153392	Malignant neoplasm of nasopharynx, unspecified	220	nasopharynx carcinoma	196	1.99E-02
C0018916	Hemangioma of unspecified site	372	capillary hemangioma	9	3.92E-02
C0018916	Hemangioma of unspecified site	372	hemangioma	38	4.04E-02
C0023418	Leukemia of unspecified cell type	79	Li-Fraumeni syndrome	7	4.95E-02
C0039075	Syndactyly	121	syndactyly	7	5.03E-02

Table 3. Overlapping pair of ICD9-ICD9 codes with temporal relationships data obtained from Hanauer and Ramakrishnan [36].

First ICD9 (CUI)	Second ICD9 (CUI)	First ICD9 Description	Second ICD9 Description
C0007102	C0018916	Malignant neoplasm of colon, unspecified	Hemangioma of unspecified site
C0001308	C0008370	Acute and subacute necrosis of liver	Obstruction of bile duct
C0007102	C0004096	Malignant neoplasm of colon, unspecified	Asthma
C0007102	C0024623	Malignant neoplasm of colon, unspecified	Malignant neoplasm of stomach, unspecified
C0014868	C0007107	Esophagitis, unspecified	Malignant neoplasm of larynx, unspecified
C0007102	C0007107	Malignant neoplasm of colon, unspecified	Malignant neoplasm of larynx, unspecified
C0007102	C0014868	Malignant neoplasm of colon, unspecified	Esophagitis, unspecified
C0018916	C0006840	Hemangioma of unspecified site	Candidiasis of unspecified site
C0007102	C0006840	Malignant neoplasm of colon, unspecified	Candidiasis of unspecified site
C0018916	C0014868	Hemangioma of unspecified site	Esophagitis, unspecified
C0007107	C0024623	Malignant neoplasm of larynx, unspecified	Malignant neoplasm of stomach, unspecified
C0014868	C0024623	Esophagitis, unspecified	Malignant neoplasm of stomach, unspecified
C0004096	C0006840	Asthma	Candidiasis of unspecified site

C0024623	C0006840	Malignant neoplasm of stomach, unspecified	Candidiasis of unspecified site
C0006840	C0008370	Candidiasis of unspecified site	Obstruction of bile duct

SUPPLEMENTARY MATERIAL

Table 1. Complete paths showing the connections between ICD9 and GO terms from the sub-graph highlighted in Figure 2. In each case the path is directed and starts with ICD9 code (UMLS id) and ends with a GO term (UMLS id) and the edge type contributing to the association is shown in braces. This data is also shown in the form a network in the second tab of this excel sheet.